Title	EDR日本語辞書からの情報獲得のための概念説明文の解析
Author(s)	藤原，滋
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1265
Rights	
Description	Supervisor:奥村 学，情報科学研究科，修士

Analysis of Japanese concept explication of EDR for information acquisition from Japanese dictionary of EDR

FUJIWARA Shigeru

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 1999

Keywords: EDR electronic dictionary, EDR thesaurus, Japanese concept explication, N-gram statistics.

Recently, large scale of machine-readable dictionaries and corpora are often used in Natural Language Processing. EDR electronic dictionary is a very large scale of machine-readable database of language, that has word dictionaries and thesaurus and corpora. EDR thesaurus is a part of EDR electronic dictionary, which consists of 400,000 concepts, and concept explication explain each concept. We can get useful information from EDR thesaurus, e.g. hypernym, hyponym, synonym. For users need these information, the rich concepts of EDR thesaurus is useful.

However, for processes using EDR thesaurus, performance of EDR thesaurus is not always enough. For example,

- The distribution of concepts of EDR thesaurus is slanted.
- The number of synonym given by EDR thesaurus is too much. More closely synonyms are needed.

The former case are solved by adding new concepts by editor. However, in the latter case, the building thesaurus which covering all requirements is not realistic because a level of requirement is inconsist for every users. In this paper, we propose method of adding such information as distinguish difference among synonyms.

We extract this information from concept explication. However, information of concept explication can not use by computer because they are written in natural language. So we make morphological, syntactical, semantical analysis, for making machine handle information of concept explication.

First, we notice similarity between concept explication and explanation of ordinary dictionaries, tried to get information of syntactic structure using surface feature of concept explication. For the reason, we get N-gram statistics from set of concept explication which classified part of speech. As the result, different frequently N-grams every part of speech are acquired and it is known that some of them give information which contribute syntactical and semantical analysis. (we call such N-gram **key word**.) In this paper, we define **definite word**, **to-iu part**, **ni-oite part**. These information are extracted easily by using information of **key word**.

We explain the process of semantical analysis of concept explication. First, making groups which consist of synonyms. Secondly, making groups which consist of same word (we call such group **A-group**)and making groups which consist of word have same govern word and same case. (we call such group **B-group**.) After grouping process, do word sense disambiguation for definite words and words of to-iu part and words of ni-oite part by using **a semantical restriction given by key words** and **scoring based information of EDR thesaurus**. If a A-group have a word which fixed sense already exist, fix sense of all words of this A-group. Words of B-group is decided sense by using **information that they have similar senses with each other**. Words is not decided sense until this process is decided sense by applying default method based information of frequency of EDR Co-occurrence dictionary.

For verification of effectiveness of method is proposed us, we did experiment. We make 21 test sets which consist of leaf node of EDR thesaurus. (Sum of concepts of test set is 535.) As the result, we get good result relatively for definite words and words of to-iu part and words of ni-oite words, recall is 56% and precision is 64%. And we get result for other words by applying default method, recall is 60%, precision is 55%. Lastly, we get result for all words of test set, recall is 60%, precision is 55%.

Especially, the method for definite words and words of to-iu part and words of ni-oite part get good result however it consist of low-cost methods.

Through the experiment, accuracy is not very high, but there is still room for improvement in this score and ranking.